

Optimal Cooperative Cognitive Relaying and Spectrum Access for an Energy Harvesting Cognitive Radio: Reinforcement Learning Approach

Ahmed El Shafie[†], Tamer Khattab^{*}

[†]Wireless Intelligent Networks Center (WINC), Nile University, Giza, Egypt.

^{*}Electrical Engineering, Qatar University, Doha, Qatar.

Abstract—In this paper, we consider a cognitive setting under the context of cooperative communications, where the cognitive radio (CR) user is assumed to be a self-organized relay for the network. The CR user cooperatively relays some of the undelivered packets of the energy harvesting primary terminal. Specifically, the CR user stores and relays those packets which it receives successfully from the primary, but are not decoded correctly by the primary destination. The CR user manages the fraction of the undelivered primary packets to be accepted. Moreover, it has the decision of choosing the queue from which it will transmit at the idle time slots (slots where the primary user (PU) is inactive). The channel is assumed to be slotted and one time slot is sufficient for the transmission of one data packet. The CR user and the PU are assumed to be energy harvesters. It is assumed that one data packet transmission dissipates one energy packet. The optimal policy is varying according to the primary and CR users arrival rates as well as the channels connectivity. The CR user saves energy for the PU by taking the responsibility of relaying the undelivered primary packets. It optimally organizes its own energy packets to maximize its payoff as time progresses.

Index Terms—Cognitive radio, reinforcement learning, energy harvesting, Q -learning, optimal policy.

I. INTRODUCTION

Secondary utilization of the licensed frequency bands can efficiently improve the spectral density of the underutilized licensed spectrum. Cognitive radio (CR) users are smart terminals that use cognitive technologies to be fully aware of the environmental variations. Moreover, CR terminals should exploit methodologies of learning and reasoning to dynamically reconfigure their communication parameters.

Cooperative diversity, which is a recently emerging technique for wireless communications, has gained wide attention recently. Cooperative cognitive relaying, which involves cooperation between primary and secondary nodes, has been investigated in many works, e.g., [1]–[3]. In [1], the authors investigate a cognitive network with one primary user (PU) and one CR user. The cognitive terminal optimally adjusts its power such that the secondary queue mean service rate is maximized while maintaining all queues in the network stable. In [2], the authors considered that the CR terminal can use the primary spectrum when the PU is inactive under a priority in transmission assigned to the relaying queue. The CR user admits a predefined fraction of the undelivered packets of the PU to be relayed. The authors optimize over

that fraction to achieve the minimum secondary queueing delay. In [3], the stable-throughput region of a network with one PU and one secondary user (SU) in cognitive shared channel is characterized. The PU transmits its packet whenever it has packets. The primary transmission takes place from the beginning of the time slot till the far end of the slot if the primary queue is nonempty, whereas the cognitive user transmits its packets with probability one if the PU is inactive, and it transmits with some probability if the PU is active. The CR user can use the channel at the same time with the PU due to the multipacket reception (MPR) capability supplied to the receiving nodes.

Energy harvesting technology has been emerged recently to transmitting terminals. Optimal energy management has been addressed in many papers such as [4]–[6]. The authors of [4], Sharma et al., obtained the optimal energy management policies for an energy harvester. In [5], energy allocation over a finite horizon is considered with the objective of maximizing the throughput and taking into account time-varying channel conditions. In [6], communication by an energy harvester over a wireless fading channel is considered. Stochastic dynamic programming is used to solve for the optimal online policy that maximizes the average number of bits delivered by a deadline under stochastic fading and energy arrival processes with causal channel state feedback.

In a cognitive setting, there are several works which include energy harvesting, e.g., [7]–[13]. In [7], a Markov decision process (MDP) is proposed to obtain the optimal secondary access policy under perfect spectrum sensing. The authors of [8] investigate an energy constrained cognitive terminal without explicitly involving an energy queue. The authors of [9] investigate a scenario with one rechargeable PU and one cognitive terminal. The maximum stable throughput region is characterized. In [10], the authors investigate the maximum stable secondary mean service rate under the stability of the primary and secondary queues and with MPR capability added to the receiving nodes. The network model consists of a PU and an energy harvester CR user. In [11], Krikidis et al. investigate the impact of cooperation in a wireless three-node network with energy harvesting nodes and bursty data traffic from network layer standpoint. The authors derive the stability region of the system as well as the required transmitted power for both a non-cooperative and an orthogonal decode-and-forward cooperative protocols. In [12], the authors assume a

simple access scheme where the SU randomly accesses the channel at the beginning of the time slot without performing channel sensing to exploit the MPR capability of receiving nodes. The maximum throughput of a saturated SU is obtained under stability and queueing delay constraints on the primary queue. In [13], El Shafie et al. proposed a cognitive setting with one energy harvesting PU and one energy harvesting SU. The SU randomly selects a sensing duration from a predefined set to ascertain the primary activity. The authors obtained the maximum stable throughput of the SU under stability of the PU's queue.

In this paper, we assume that the CR user senses the channel for τ seconds from the beginning of the time slot. Based on the sensed primary state, the SU has to take an action. Thus, the action is taken after τ seconds from the beginning of the time slot; exactly after sensing the channel. If the PU is sensed to be inactive, the CR user has to choose between being idle to the end of the time slot or to transmit a packet either from its own packets or from the relaying packets. If the PU is active, the CR user has to choose between being idle or accepting a primary packet. Unlike most of the existing works, we do not assume M/D/1 with unity service rate model for the energy queues (for details, the reader is referred to [9], [10] and the references therein), which is a trivial model and it provides an inner bound on the performance and makes the queue capacity useless as shown in [14]. Moreover, we assume a finite length energy and data queues. We do not consider either dominance system approach or always nonempty queues to decouple the queues as proposed in many works [9], [10], [15]. Furthermore, in contrast to the conventional modeling of the arrival processes as Bernoulli identically and independent random processes [9], [10], we consider correlated arrivals at each queue and modeling the arrival processes of the queues as *Markov modulated Bernoulli processes*. The proposed approach and the analysis presented in this paper are generic and can be applied to any system.

This paper is organized as follows. In Section II, we describe the system model adopted in this paper. The service processes of the queues are discussed in Section III. In Section IV, we present the reinforcement formulation and *Q-learning* algorithm. We describe the simulation scenario considered for performance evaluation in Section V. Finally, we conclude the paper in Section VI.

II. SYSTEM MODEL

The network model adopted in this paper composes of two energy harvesters sharing the same channel resource and with different priorities. The PU is the terminal with the highest priority and it has unconditional access to the channel, whereas the CR user is the lowest priority terminal and it accesses the channel whenever the primary terminal is declared to be inactive. The inactivity of the PU can be due to the lack of energy packets at its energy queue or due to the lack of data packets at its data queue. The network is depicted in Fig. 1.

We assume that the primary transmitter has two different types of buffers; a data buffer to store its incoming data packets, denoted as Q_p , and an energy buffer to store the

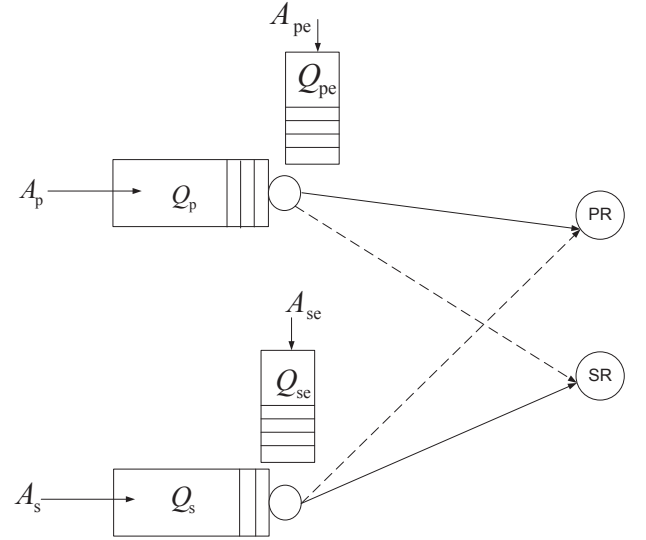


Fig. 1. Primary and secondary links and queues. The solid lines are the communication channels, while the dotted lines from the receivers to the transmitters are the feedback and the interference channels. The primary and secondary receivers are denoted by PR and SR, respectively.

energy harvested from the environment, denoted as Q_{pe} . The CR user has three buffers; Q_s to store its own arrived data traffic, Q_{ps} to store the accepted for relaying undelivered packets of the PU in case of outage on the primary channel, and finally, Q_{se} to store the harvested energy packets from the environment. We assume all buffers are of **finite** length. Precisely, queue Q_j , $j \in \{p, pe, s, ps, se\}$, has at most B_j packets. We consider a time-slotted transmission where all packets have the same size and one time slot is exactly suits the transmission of one data packet. The packets arrival processes of PU and SU queues are assumed to be Markov modulated Bernoulli processes [16] where the probability of arrival occurrence of a Bernoulli process evolves over time according to a Markov chain. The arrivals at each queue are assumed to respect the following two state Markov chain (shown in Fig. 2):

$$\begin{pmatrix} 1 - \lambda_k & \lambda_k \\ \beta_k & 1 - \beta_k \end{pmatrix}$$

where λ_k denotes the probability of having no arrived packet at queue Q_k , $k \in \{p, pe, s, se\}$, in time slot $t+1$ when there was no arrived packet in time slot t and β_k denotes the probability of having no arrived packet at queue Q_k in time slot $t+1$ when there was an arrived packet in time slot t . We assume that arrivals are independent random variables from queue to queue. We denote the arrival at Q_k by A_k^t where $A_k^t = 1$ if there is an arrived packet in time slot t and zero otherwise.

The radio channel gain of the channels between any pair of nodes h_i^t is assumed to be zero mean circularly symmetric complex Gaussian random variable with variance σ_i^2 i.e. $\mathcal{CN}(0, \sigma_i^2)$ and independent for all i , where i reads 'p' for the primary link, 's' for the secondary link, 'ps' for the link between the PU and the CR user and 'sp' for the link between the CR user and the primary destination. Each link is perturbed

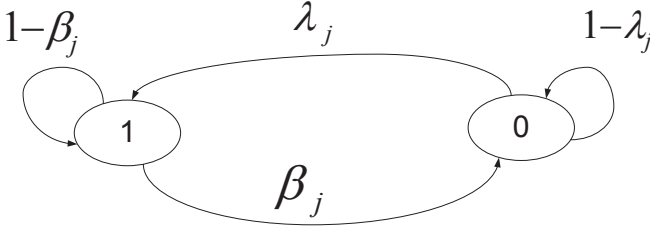


Fig. 2. Two state Markov model of Markov modulated Bernoulli process for queue Q_k , $k \in \{p, pe, s, se\}$.

thermal noise which is modeled as complex additive white Gaussian noise (AWGN) with zero mean and variance \mathcal{N}_o and independent for all links. We assume a two state Markov channels. Specifically, the i th link follows (shown in Fig. 3):

$$\begin{pmatrix} 1 - \Gamma_i & \Gamma_i \\ q_i & 1 - q_i \end{pmatrix}$$

where Γ_i is the probability of link i being not in outage at time $t + 1$ given that it was in outage at time t and q_i is the probability of the link being in outage in time $t + 1$ given that it was not in outage at time t .

Let $\overline{\mathcal{X}} = 1 - \mathcal{X}$, $1[\mathcal{F}] = 1$ if the argument event \mathcal{F} is true, and $I_{c_i}^t$ be the indication of the channel state and it is equal to unity if link i is connected and zero otherwise. We consider that the channel is ON (connected) if the transmitted rate is less than or equals to the channel capacity; otherwise, it is OFF (disconnected). We assume that the SU knows the channels gains perfectly at the beginning of the time slot. The primary channel can be sent from the primary destination over a dedicated narrow band during the sensing phase of the spectrum.¹ We define $\overline{I}_{c_i}^t = 1 - I_{c_i}^t$ as the state of connectivity of the channel i in time slot t .

The medium access control is assumed to obey the following rules.

- At the beginning of the time slot, the CR user senses the channel for τ seconds from the beginning of the time slot to declare the state of activity of the PU.²
- The sensing process result is recorded as a binary value at the secondary terminal. In particular, it is recorded as '1' if the PU is active or '0' if the PU is inactive.
- If the PU is sensed to be inactive, the CR user has to choose between being idle till the end of the time slot or to transmit a packet either from its own queue, Q_s , or from the relaying queue, Q_{ps} .
- If the PU is active, the CR user has to choose between being idle till the end of the time slot or to accept the primary packet.
- If the primary receiver could not decode the PU packet correctly and the CR user decided to accept the packet at the beginning of the time slot, it has to send acknowledgement/negative-acknowledgment

¹We would emphasize that the proposed protocol is based on the cooperation between users. Thus, the primary system aids the secondary system for increasing their performance at the same time.

²The sensing duration should be large enough for channels status estimation and perfect channel sensing.

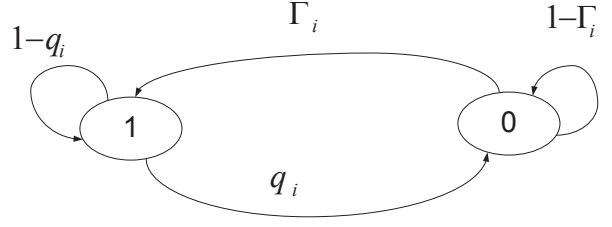


Fig. 3. Two state Markov model for link i .

(ACK/NACK) messages to the PU based on the result decoding of the packet. These packets are then dropped from the primary queue.

- In case both the CR user and the primary destination fail to decode the primary data, a retransmission of the packet is initiated by the PU at the following time slots.

We assume that the overhead for transmitting the ACKs and NACKs is negligible relative to packet sizes. The second assumption we make is that the errors in packet acknowledgement feedback is negligible, which is reasonable for short length ACK/NACK packets as low rate codes can be employed in the feedback channel [15]. In addition, nodes cannot transmit and receive at the same time which is a common assumption where terminals are equipped with single transceivers [3].

According to the previous description, the SU has four distinct actions. After τ second from the beginning of the time slot, it has to choose one of the possible actions. The CR user should optimally distribute its energy packets among the transmissions of the data packets to achieve the highest possible performance.

III. QUEUES ARRIVAL AND SERVICE PROCESSES

As mentioned earlier, the CR user has four possible actions. The set of actions is $\mathcal{A} = \{a_1, a_2, a_3, a_4\}$, where a_1 : transmitting a packet from Q_s , a_2 : transmitting a packet from Q_{ps} , a_3 : accepting a packet from the PU, and a_4 : remaining idle (CR user is idle). Note that the optimal action vector in a given time slot satisfies the constraint:

$$\sum_{i=1}^4 a_i^t = 1, \quad \forall t = 0, T, 2T, \dots \quad (1)$$

This condition means that there is only one action per time slot.

A packet from Q_s is served if the CR user has energy packets in its energy queue, the CR user accesses the channel using Q_s , the channel to the respective receiver is ON, and the PU is inactive. Mathematically, the service process can be modeled as:

$$\mathcal{R}_s^t = a_1 I_{c_s}^t \left(1 - I_{Q_p}^t I_{Q_{pe}}^t \right) I_{Q_{se}}^t \quad (2)$$

The term $I_{Q_j}^t$ equals to unity if the queue Q_j is not empty and zero if the queue is empty. Note that the PU is active if both its data and energy queues are nonempty, i.e., $I_{Q_p}^t I_{Q_{pe}}^t = 1$. Thus, the term $1 - I_{Q_p}^t I_{Q_{pe}}^t$ indicates the inactivity of the PU.

Now, we move our attention to the relaying queue. A packet from Q_{ps} is served if the CR user has energy in its energy queue, the CR user decides to access the channel with a packet from Q_{ps} , the channel to the respective receiver is ON, and the PU is inactive. Mathematically, the service process can be modeled as:

$$\mathcal{R}_{ps}^t = a_2 I_{c_{sp}}^t \left(1 - I_{Q_p}^t I_{Q_{pe}}^t\right) I_{Q_{se}}^t \quad (3)$$

The arrival process to the relaying queue is described as follows. A packet is arrived to the relaying queue if the primary queue is nonempty, the relaying queue is not full, the channel between the PU and its respective receiver is OFF, the channel from the PU to the secondary user is ON, and the SU decides to accept the packet. Mathematically, the arrival process is given by

$$A_{ps}^t = a_3 I_{c_{ps}}^t I_{Q_p}^t I_{Q_{pe}}^t \overline{I_{c_p}^t} 1[Q_{ps} < B_{ps}] \quad (4)$$

A packet from the secondary energy queue is consumed in either one of the following events. If the SU accesses the channel either from its data queue or from the relaying queue. Mathematically, the process is given by

$$\mathcal{R}_{se}^t = a_1 I_{Q_s}^t + a_2 I_{Q_{ps}}^t \quad (5)$$

Given that the PU has energy in its energy queue, a packet from the PU's data queue is served in either one of the following events. If the PU channel to its respective receiver is ON and the SU remains idle; or if the channel between the PU and PD is OFF, the channel between the PU and the SU is ON, the relaying queue is not full, and the SU decides to accept the packet. The process is modeled as follows:

$$\mathcal{R}_p^t = I_{Q_{pe}}^t \left(a_4 I_{c_p}^t + a_3 I_{c_{ps}}^t \overline{I_{c_p}^t} 1[Q_{ps} < B_{ps}] \right) \quad (6)$$

Since the PU accesses the channel unconditionally whenever it has energy and data packets, a packet from the PU energy queue is consumed if the primary queue is nonempty. That is,

$$\mathcal{R}_{pe}^t = I_{Q_p}^t \quad (7)$$

We assume that departures occur before arrivals, and the queue size is measured at the beginning of the slot [17]. The evolution of queue Q_j is given by

$$Q_j^{t+1} = \min \left\{ \max \left\{ Q_j^t - \mathcal{R}_j^t, 0 \right\} + A_j^t, B_j \right\}, j \in \{p, pe, s, ps, se\} \quad (8)$$

where $\max\{.,.\}$ and $\min\{.,.\}$ return the maximum and the minimum among the values in the argument, respectively.

IV. Q-LEARNING ALGORITHM

The prime goal in the reinforcement learning (RL) is to choose actions over time so as to maximize the expected value of the total payoff of the learner (agent or user). The CR user will be able to achieve the adaptive optimal policy according to the mean arrival rates of the queues and outage probabilities of all channels in order to maximize its expected payoff as time progresses. MDPs are considered as a powerful

framework for solving problems of sequential decision making under uncertainty [18]–[20]. Bellman's equation, which forms the foundation for many dynamic programming approaches to solving MDPs, is given by:

$$V(s) = \mathcal{R}(s, a) + \gamma \sum_{\hat{s} \in \mathcal{S}} P(\hat{s}|s, a) V(\hat{s}) \quad (9)$$

where $V(s)$ is the discounted cumulative reward and γ is a constant that determines the relative value of delayed versus immediate rewards. Choosing the discount factor γ smaller than 1 ensures convergence of the sum. For every state s we may investigate what the best policy (action) is, and what its value would be. Let us define the optimal value function as the maximum value function among all value functions, it satisfies the Bellman equation, and is given by

$$V^*(s) = \max_a \left[\mathcal{R}(s, a) + \gamma \sum_{\hat{s} \in \mathcal{S}} P(\hat{s}|s, a) V(\hat{s}) \right] \quad (10)$$

where $V^*(s)$ gives the maximum discounted cumulative reward that the agent can obtain starting from state s ; that is, the discounted cumulative reward obtained by following the optimal policy beginning at state s [18]. The policy is a function that maps the state space to action space, i.e., $\pi : \mathcal{S} \rightarrow \mathcal{A}$. The optimal policy is given by:

$$\pi^*(s) = \operatorname{argmax}_a \left[\mathcal{R}(s, a) + \gamma \sum_{\hat{s} \in \mathcal{S}} P(\hat{s}|s, a) V(\hat{s}) \right] \quad (11)$$

The reward function is defined according to the states and actions and it aims at maximizing the weighted sum of the throughput of the CR user transmitter's queue and the relaying queue subject to some predefined constraints. Mathematically, the immediate reward function is given by:

$$\begin{aligned} \mathcal{R}(s, a) = & \omega \mathcal{R}_s I_{Q_s} + (1 - \omega) \mathcal{R}_{ps} I_{Q_{ps}} \\ & - \mathcal{K} \left[I_{Q_p} I_{Q_{pe}} (a_1 + a_2) + a_1 \overline{I_{c_s}} \overline{I_{Q_s}} \overline{I_{Q_{se}}} \right. \\ & + a_2 \overline{I_{c_{ps}}} \overline{I_{Q_{ps}}} \overline{I_{Q_{se}}} + a_3 1[Q_{ps} = B_{ps}] \\ & \left. + a_3 (I_{c_p} I_{Q_p} I_{Q_{pe}} + \overline{I_{c_{ps}}} \overline{I_{Q_p}} \overline{I_{Q_{pe}}}) \right] \end{aligned} \quad (12)$$

where ω is a fixed constant that belongs to the set $[0, 1]$ and \mathcal{K} is a penalty constant. The rational behind this cost function is that the CR users cannot transmit at the same time with the PU to avoid a sure collision event, which is specified by $-\mathcal{K} I_{Q_p} I_{Q_{pe}} (a_1 + a_2)$; to avoid wasting secondary energy when channels are in outage, which is specified by $-\mathcal{K} (a_1 (\overline{I_{c_s}} \overline{I_{Q_s}} \overline{I_{Q_{se}}}) + a_2 \overline{I_{c_{ps}}} \overline{I_{Q_{ps}}} \overline{I_{Q_{se}}})$; to avoid decoding the primary packet when the relaying queue is full, which is specified by $a_3 1[Q_{ps} = B_{ps}]$; and to avoid using an action when the corresponding queue is empty or the secondary energy queue is empty or to take packet acceptance action when the PU is inactive. Note that, the more ω indicates more emphasizing on the service rate of Q_s (secondary throughput), and the lower the ω the more emphasizing on the service rate of Q_{ps} .

***Q-learning* algorithm**

```

Initialize:
let  $t = 0$ 
for each  $s \in \mathcal{S}$  and  $a \in \mathcal{A}$  do
  initialize the  $Q$  value
end for
Initialize  $s^t$ 
Learning:
loop
  generate a random number  $\ell$  between 0 and 1
  if  $\ell < \mu$  then
    select action randomly
  else
    select the action  $a^t$  characterized by the
    maximum  $Q$ -value
  end if
  execute  $a^t$ 
  receive an immediate reward  $\mathcal{R}(s^t, a^t)$ 
  observe the next state  $s^{t+1}$ 
  update the table entry as follows:
   $s^t \leftarrow s^{t+1}$ 
   $Q(s, a) \leftarrow Q(s, a) + \alpha \left( \mathcal{R}(s, a) + \gamma \max_{\hat{a}} Q(\hat{s}, \hat{a}) - Q(s, a) \right)$ 
end loop

```

In *Q-learning*, the agent, which is the CR user in this work, interacts with the environment to obtain the consecutive actions that maximize the accumulative payoff of the weighted sum of the secondary queues, Q_s and Q_{ps} , mean service rates. In particular, the CR user aims at maximizing the expected weighted sum of the its queue service rates. It is assumed that the environment is a finite-state discrete time stochastic dynamical system.

The interactions between the CR user and the environment at every time instant t is described as follows.

- The CR user senses the channel for τ seconds.
- The CR user observes its state s .
- Based on s , the CR user chooses an action a from the feasible actions set \mathcal{A} .
- The CR user receives an immediate reward $\mathcal{R}(s, a)$.
- A transition to the state \hat{s} takes place.
- The learning process is repeated until convergence to the optimal policy.

The *Q-learning* algorithm (described at the top of this page) is the most popular powerful and widely used form of reinforcement learning due to the naive of implementation of this method. It obtains the optimal *Q-values*, rather than state-values. The update rule for *Q-learning* is

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[\mathcal{R}(s, a) + \gamma \max_{\hat{a}} Q(\hat{s}, \hat{a}) - Q(s, a) \right] \quad (13)$$

where α is the learning rate and γ is the discount factor. The idea of update rule is that the part $\mathcal{R}(s, a) + \gamma \max_{\hat{a}} Q(\hat{s}, \hat{a})$ is an estimate of the *Q-value* $Q(s, a)$. Watkins proved that this method will converge to the *Q-values* for the optimal policy, $Q^*(s, a)$, if two conditions were met, every state-action pair

has to be visited infinitely often and learning rate α decay over time. A proof of convergence for *Q-learning* based on that outlined in Watkins was presented in [21]. They show that *Q-learning* converges to the optimum action-values with probability 1 so long as all actions are repeatedly sampled in all states and the action-values are represented discretely. The objective of the CR user is to find an optimal policy $\pi^*(s) \in \mathcal{A}$ for each state s , to maximize some cumulative measure of the cost $\mathcal{R}(s, a)$ received over time. We define the evaluation function, denoted by $Q(s, a)$, as the expected total discount cost over an infinite time and it is given by

$$Q(s, a) = \mathcal{E} \left\{ \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s, \pi(s)) | s_0 = s \right\} \quad (14)$$

where $\mathcal{E}\{\cdot\}$ denotes the expected value. If the selected action a at time t following the policy $\pi(s)$ which is corresponding to the optimal policy $\pi^*(s)$, the *Q-function* is maximized with respect to the current state. It can be shown that (14) is given by

$$Q(s, a) = \mathcal{E} \left\{ \mathcal{R}(s, a) \right\} + \gamma \sum_{\hat{s} \in \mathcal{S}} P(\hat{s} | s, a) Q(\hat{s}, a) \quad (15)$$

Recall that $P(\hat{s} | s, a)$ is the transition probability from state s to next state \hat{s} , when action a is executed. Eqn. (15) indicates that the *Q-function* of the current state-action pair, can be represented in terms of the expected immediate cost of the current state-action pair and the *Q-value* of the next state-action pairs. *Q-learning* aims at determining an optimal stationary policy $\pi(s)$, without knowing $\mathcal{E}\{\mathcal{R}(s, a)\}$ and $P(\hat{s} | s, a)$. The states are defined as follows. Without loss of generality, we divide the CR user's queues to \mathcal{N} portions. In particular, each queue in the CR terminal is divided to \mathcal{N} portions as follows:

$$\mathcal{L}(Q_n) = \begin{cases} 0 & \text{if } Q_n = 0 \\ 1 & \text{if } 0 < Q_n \leq \nu_{n,th,1} \\ 2 & \text{if } \nu_{n,th,1} + 1 \leq Q_n \leq \nu_{n,th,2} \\ 3 & \text{if } \nu_{n,th,2} + 1 \leq Q_n \leq \nu_{n,th,3} \\ \vdots & \vdots \\ \mathcal{N}-1 & \text{if } Q_n \geq \nu_{n,th,\mathcal{N}-2} + 1 \end{cases} \quad (16)$$

where $n \in \{s, ps, se\}$ and $\nu_{n,th,h}$ is the h th threshold of the queue Q_n .

The state vector, at any time instant t , is formed as

$$\mathcal{S}^t = \left[I_{Q_p}^t, I_{Q_{pe}}^t, \mathcal{L}(Q_{ps}^t), \mathcal{L}(Q_{se}^t), \mathcal{L}(Q_s^t), I_{c,sp}^t, I_{c,s}^t, I_{c,p}^t, I_{c,ps}^t \right] \quad (17)$$

where $I_{Q_p}^t, I_{Q_{pe}}^t$ represents the activity of the PU and is ascertained from channel sensing. According to the above description, the total number of states is $2^5 \times \mathcal{N}^3$, where 2 represents the possibility of the binary valued channels states.

With respect to the *Q-learning* algorithm, the learning rate is $\alpha = 0.5$ and the discount factor is $\gamma = 0.9$. We also introduce a probability $\mu = 0.05$ of visiting random states in the initial 60% of the *Q-learning* iterations. This parameter is used in the action selection procedure to guarantee that the final policy is a global optimum and not a local one [22].

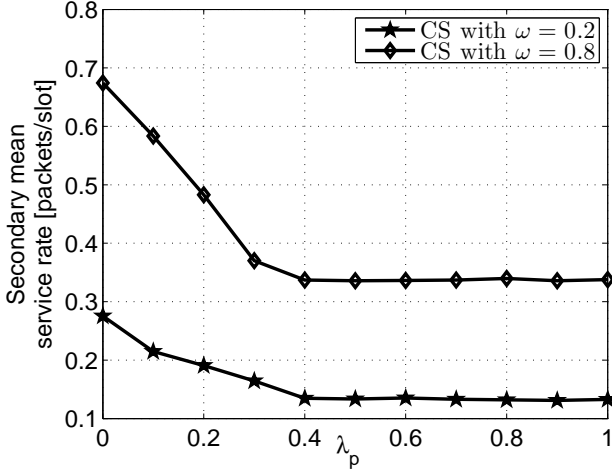


Fig. 5. The maximum secondary service rate in packets/slot.

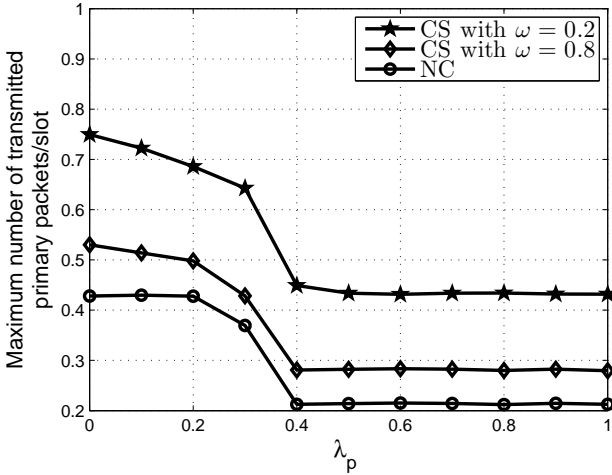


Fig. 4. The maximum primary transmitted packets per time slot.

V. RESULTS AND SIMULATIONS

In this section, we provide some simulations of the system. Simulations are executed using $\beta_p = \lambda_p \in [0, 1]$. Let ‘CS’ denote the cooperative system and ‘NC’ denote the non-cooperative system. We assume that each buffer, in the network, of size 20 packets. We split each queue of the CR user to $\mathcal{N} = 4$ portions which leads to the availability of $\mathcal{N}^3 = 64$ states. The thresholds of the queues are: $\nu_{n,th,0} = 6$ and $\nu_{n,th,1} = 12$ for all n . The capacity of all queues belonging to the system is $\mathcal{B}_j = \mathcal{B} = 20$, for all j , packets. The rest of the parameters are: $\mathcal{K} = 10$, $\Gamma_p = 0.2$, $q_p = 0.4$, $\Gamma_s = 0.6$, $q_s = 0.1$, $\Gamma_{ps} = 0.7$, $q_{ps} = 0.2$, $\Gamma_{sp} = 0.8$, $q_{sp} = 0.05$, $\lambda_s = 0.4$, $\beta_s = 0.4$, $\lambda_{se} = 0.8$, $\beta_{se} = 0.4$, $\lambda_{pe} = 0.4$, $\beta_{pe} = 0.4$ and $\beta_p = \lambda_p$.

As shown in Fig. 4, the primary maximum number of transmitted packets per time slot increases with cooperation. Moreover, decreasing ω increases the service rate of the relaying queue; hence, increases the primary transmitted packets per time slot. The beneficial gain of cooperation is shown in the figure. Fig. 5 demonstrates the maximum mean service rate

of the secondary own data queue. As seen from the figure, increasing ω emphasizes on the secondary service; hence, increases the secondary mean service rate. From the figures, we conclude that the cooperation is important for both users and it boosts their throughputs. Furthermore, The parameter ω manages the throughputs of users and it can be used to archive certain quality of service for each user.

VI. CONCLUSION

In the paper, we have investigated a cooperative energy harvesting CR user sharing the channel resources with a PU. The CR user has to decide on taking a specific action preceded by a channel sensing from the the beginning of every time slot. The action taken at each state is selected to, on the average, maximize the secondary expected utility as time progresses. Unlike most of the existing work, we have considered finite queue lengths and characterized the system performance with the existence of strong queue interaction. We also have considered Markov modulated Bernoulli arrival processes at queues. The optimal policy has been obtained using *Q-learning* algorithm where each state is assigned an action.

REFERENCES

- [1] O. Simeone, Y. Bar-Ness, and U. Spagnolini, “Stability analysis of the cognitive interference channel,” in *Fortieth Asilomar Conference on Signals, Systems and Computers*, 29 2006–Nov. 1 2006, pp. 1357–1361.
- [2] M. Elsaadany, M. Abdallah, T. Khattab, M. Khairy, and M. Hasna, “Cognitive relaying in wireless sensor networks: Performance analysis and optimization,” in *Proc. IEEE GLOBECOM*, Dec. 2010, pp. 1–6.
- [3] S. Kompella, G. Nguyen, J. Wieselthier, and A. Ephremides, “Stable throughput tradeoffs in cognitive shared channels with cooperative relaying,” in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 1961–1969.
- [4] V. Sharma, U. Mukherji, V. Joseph, and S. Gupta, “Optimal energy management policies for energy harvesting sensor nodes,” *IEEE Trans. Wireless Commun.*, vol. 9, no. 4, pp. 1326–1336, 2010.
- [5] C. Ho and R. Zhang, “Optimal energy allocation for wireless communications powered by energy harvesters,” in *Proc. IEEE ISIT*, 2010, pp. 2368–2372.
- [6] O. Ozel, K. Tutuncuoglu, J. Yang, S. Ulukus, and A. Yener, “Transmission with energy harvesting nodes in fading wireless channels: Optimal policies,” *J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1732–1743, 2011.
- [7] S. Park, S. Lee, B. Kim, D. Hong, and J. Lee, “Energy-efficient opportunistic spectrum access in cognitive radio networks with energy harvesting,” in *Proceedings of the 4th International Conference on Cognitive Radio and Advanced Spectrum Management*. ACM, 2011, p. 62.
- [8] A. Hoang, Y. Liang, D. Wong, Y. Zeng, and R. Zhang, “Opportunistic spectrum access for energy-constrained cognitive radios,” *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1206–1211, 2009.
- [9] N. Pappas, J. Jeon, A. Ephremides, and A. Traganitis, “Optimal utilization of a cognitive shared channel with a rechargeable primary source node,” *JCN*, vol. 14, no. 2, pp. 162–168, 2012.
- [10] A. El Shafie and A. Sultan, “Optimal random access and random spectrum sensing for an energy harvesting cognitive radio,” in *Proc. IEEE WiMob*, Barcelona, Spain, Oct. 2012, pp. 403–410.
- [11] I. Krikidis, T. Charalambous, and J. Thompson, “Stability analysis and power optimization for energy harvesting cooperative networks,” *IEEE Signal Processing Letters*, vol. 19, no. 1, pp. 20–23, 2012.
- [12] A. El Shafie and A. Sultan, “Optimal random access for a cognitive radio terminal with energy harvesting capability,” *IEEE Commun. Lett.*, vol. 17, no. 6, pp. 1128–1131, 2013.
- [13] —, “Optimal selection of spectrum sensing duration for an energy harvesting cognitive radio,” in *Proc. IEEE GLOBECOM*, Dec 2013, pp. 1020–1025.
- [14] —, “Comments on ‘Optimal Utilization of a Cognitive Shared Channel with a Rechargeable Primary Source Node’.” Available [Online]:<http://arxiv.org/pdf/1401.3174v1.pdf>, 2014.

- [15] A. Sadek, K. Liu, and A. Ephremides, "Cognitive multiple access via cooperation: protocol design and performance analysis," *IEEE Trans. Info. Theory*, vol. 53, no. 10, pp. 3677–3696, Oct. 2007.
- [16] S. Özekici, "Markov modulated bernoulli process," *Mathematical Methods of Operations Research*, vol. 45, no. 3, pp. 311–324, 1997.
- [17] R. Rao and A. Ephremides, "On the stability of interacting queues in a multiple-access system," *IEEE Trans. Info. Theory*, vol. 34, no. 5, pp. 918–930, Sep. 1988.
- [18] T. Mitchell, *Machine learning*. Burr Ridge, IL: McGraw Hill, 1997.
- [19] R. Sutton and A. Barto, *Reinforcement learning: An introduction*. Cambridge University Press, 1998, vol. 28.
- [20] R. Sutton, D. Precup, and S. Singh, "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning," *Artificial intelligence*, vol. 112, no. 1, pp. 181–211, 1999.
- [21] C. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3, pp. 279–292, 1992.
- [22] A. Galindo-Serrano and L. Giupponi, "Distributed Q-learning for aggregated interference control in cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 4, pp. 1823–1834, May 2010.